Leah Aguilar

Dr. Don Berkich

PSYC-4390-003

4/27/23

# AI’s Potential to Induce Mental Atrophy

## Introduction:

In most major discussions, being held on whether ‘Artificial Intelligence’ (‘AI’) will eventually advance to the point to where it can withstand logical scrutiny to be classified as a true synthetic mind comparable to human capabilities. If not for that argument, it is being discussed on when AI will replace jobs traditionally held by humans. Neither discussion seems very fruitful to deliberate at this point since the answer for both remains obvious --- if there is a profit motive for owners of AI technology to implement their machines.

An argument on AI that ought to be made, at least in higher frequencies since they make up a minority of the conversation surrounding AI, is whether it is beneficial for humanity to continue this seemingly endless road towards complete automation in the moral and the selfish business productivity sense. It is imperative that the argument against over reliance or even prioritizing AI products over human ones not just be that it would decrease public happiness to a substantial degree (though that is important to point out), but that it would lead to the eventual mental atrophy of humanity.

As it stands right now, while AI continues to advance and replace human workers, it is more productive for students of all age ranges to use AI to complete their homework. If a student would defer their work to an AI program, is that child keeping enough the basic concepts from the AI to do the work themselves? In other words, do they know how to construct an equation on their own? If they do not, then how can they understand the concepts they are being taught in school? By comparing the consequences of the reliance on graphing calculators on learners of all skill levels (with a focus on novices), it will be easier to postulate a potential future where the use of AI systems to supplement foundational or intermediate learning in k-12 and college students.

## Artificial Intelligence—What is it?

Defining what Artificial Intelligence (AI) is, especially in relation to a sapient mind, is paramount to understanding any potential capabilities it might have or will be likely to gain. According to Russell and Norvig (1995, 2002, 2009) series of papers, AI refers to the development of computer systems that can perform tasks that require human intelligence, such as visual perception, speech recognition, decision-making, and natural language processing. AI can be thought of as a field of study that focuses on creating intelligent machines that can learn, reason, and solve problems in ways similar to humans.

AI systems can be classified into two broad categories: narrow or weak AI and general or strong AI. Weak AI refers to systems that perform specific tasks, such as recognizing faces or playing chess either because of following the strict computations of a rules-based game like chess (where the system is only concerned with the moves made rather than the player themselves) or making precise approximations based on mathematical computations of human facial structure. In contrast, Artificial General Intelligence (AGI) aims to develop machines that can perform any intellectual task at a level or higher than human can perform, most notably expressing creativity and empathy.

Often, creativity is defined by the ability to create new and unique ideas. However, if creativity could only be within the parameters of being wholly unique and new, humans ought to have stopped claiming to possess this skill during the

The development of AI has been made possible by advances in computer science and machine learning, which is a subfield of AI that involves the use of statistical algorithms to enable machines to learn from data. Machine learning algorithms can be broadly classified into supervised, unsupervised, and reinforcement learning.

Supervised learning involves training an algorithm on a labeled dataset, where the input data is accompanied by a desired output. The algorithm learns to make predictions based on the input data by minimizing the difference between its predicted output and the desired output. Unsupervised learning involves training an algorithm on an unlabeled dataset, where the algorithm must learn to identify patterns or relationships in the data. Reinforcement learning involves training an algorithm to decide based on feedback from its environment.

AI has the potential to revolutionize many industries, including healthcare, finance, and transportation. However, there are concerns about its impact on employment and the potential for AI systems to exhibit bias or make unethical decisions. Therefore, it is crucial for researchers, policymakers, and industry leaders to consider the ethical and social implications of AI and ensure that these technologies are developed and used responsibly. Additionally, creativity is not solely defined by the ability to generate wholly unique ideas, and this should be considered when discussing the development of AGI.

A common argument against the development of AGI is that it is impossible to program human values into a machine. The fear is that an AGI system could make decisions that are harmful to humans because it lacks empathy or an understanding of human values. While it is true that it is challenging to program human values into a machine, it is not impossible. Researchers are already working on developing ethical frameworks for AI, and these frameworks will probably be integrated into AGI systems as well.

Another concern is the potential for AGI to be biased or discriminatory. There is a fear that AGI could perpetuate and even amplify existing biases in society. While this is a valid concern, it's important to remember that AGI is only as biased as the data it is trained on. By ensuring that AGI is trained on diverse and unbiased data, we can reduce the risk of bias and discrimination.

## AI Systems: ChatGPT

How do systems like ChatGPT function? As an AI language model, ChatGPT operates by analyzing large amounts of text data, using natural language processing (NLP) algorithms to learn patterns and relationships between words, phrases, and sentences. Natural Language Processing (NLP) algorithms are a set of computational techniques used to analyze, understand, and generate human language. These algorithms enable computers to interact with humans using natural language, such as English, instead of requiring humans to use computer programming languages or commands.

NLP algorithms function by breaking down human language into its component parts, such as words, phrases, and sentences, and then using statistical models and machine learning techniques to analyze the relationships between those parts.

Specifically, ChatGPT is based on the GPT-3.5 architecture, which is a variant of the popular GPT (Generative Pre-trained Transformer) architecture developed by OpenAI.

ChatGPT's system works by taking in a prompt or input from a user, such as a question or a statement, and then generating a response based on its analysis of the input and its database of pre-existing text data. This database is pre-trained on a massive amount of text data, including books, articles, websites, and other sources, using unsupervised learning techniques. This means that the system can analyze and understand natural language patterns and relationships without explicit human guidance.

ChatGPT’s data comes from a variety of sources, including the internet, books, articles, and other written materials. However, it’s worth noting that the system generates responses based on the statistical patterns it has learned from this data, rather than on any single or narrow group of humans’ knowledge or expertise. Furthermore, the system is frequently updated and filtered by moderators to improve its accuracy and to incorporate new data sources as they become available.

In terms of its capabilities, ChatGPT has been shown to outperform other language models on a variety of language tasks, including question-answering, summarization, and machine translation. Additionally, it has been trained on a vast amount of diverse text data, making it highly versatile and able to generate responses on a wide range of topics and domains.

That being said, there are other similar language models that have also been developed and trained on large amounts of data, such as BERT (Bidirectional Encoder Representations from Transformers) and T5 (Text-to-Text Transfer Transformer). Each of these models has its own strengths and weaknesses, and the choice of which model to use for a particular task depends on various factors, including the size and quality of the available training data, the specific task requirements, and the computational resources available.

Bidirectional Encoder Representations from Transformers (BERT) and Text-to-Text Transfer Transformer (T5) are two other variants of the Generative Pre-trained Transformer

(GPT) architecture, which ChatGPT is also based on. While all three models share some similarities, they have some key differences in their architecture and training objectives.

BERT is a language model that is trained in a bidirectional manner, meaning that it processes both the left and right context of a input sequence. This allows BERT to have a deeper understanding of the meaning of a sentence and the relationships between its constituent words. BERT is trained in two tasks: Masked Language Modeling (MLM), which involves predicting missing words in a sentence, and Next Sentence Prediction (NSP), which involves predicting whether two sentences are consecutive in prompted texts. BERT is typically used for tasks such as sentiment analysis, question-answering, and text classification.

T5, on the other hand, is a text-to-text transfer transformer that is trained on a diverse set of tasks, ranging from text classification and summarization to machine translation and question-answering. Instead of training on specific tasks like BERT, T5 is trained on a large corpus of text data and is able to generate a wide range of outputs in response to a given input prompt. T5's name comes from its training objective, which involves converting input text to output text in a variety of ways, such as translation, summarization, or question-answering.

In contrast, ChatGPT is a generative language model that is pre-trained on a large amount of diverse text data using unsupervised learning techniques. It is designed to generate coherent and contextually appropriate responses to natural language prompts, rather than being trained on specific tasks like BERT or T5.

The SuperGLUE, then, is a collection of difficult natural language understanding tasks that require complex reasoning and inference. ChatGPT-3 achieved state-of-the-art performance on this benchmark, outperforming all other models, including human performance on some tasks.

ChatGPT-4 is currently being tested as frequently as possible to get the most completed results to come to the fullest conclusion on the nature of ChatGPT's performance.

The system has also been used for building dialogue systems, where it generates natural-sounding responses to user input. In a study conducted by Google researchers, ChatGPT-2 outperformed other models on a dialogue generation task based on the Persona-Chat dataset.

OpenAI has conducted various evaluations of ChatGPT's language model in terms of accuracy and consistency, as well as other metrics such as perplexity and diversity. These evaluations are typically performed using standardized benchmarks and datasets, such as the Common Crawl and WebText datasets.

Besides OpenAI's internal evaluations, there have been several third-party evaluations of ChatGPT's language model. For example, the Stanford Question Answering Dataset (SQuAD) leaderboard has included several submissions from ChatGPT-based models, which have achieved state-of-the-art results on the task of question answering.

ChatGPT has been used in various real-world applications, such as chatbots and language translation systems, which further show its accuracy and consistency in generating natural language outputs.

ChatGPT has demonstrated impressive performance on various language tasks, it is not perfect despite the fact that, as of writing this paper, it produces errors or biases in certain contexts. Ongoing research and development are focused on improving the model's performance and addressing any remaining limitations. While the promises laid out by OpenAI's ChatGPT are seemingly comprehensive, next to Microsoft's statements, it very much looks like a lot of pretty words to cover up inaction. Mircosoft being the largest stake holder in OpenAI.

## Necessity of AI

When is AI necessary? Certainly, we ought not require NASA scientists to write out equations by hand during every single calculation, including during high stress scenarios. As it stands, some form of AI (largely weak AI) is used in a variety of contemporary life.

One of the main reasons AI is necessary is because it can help solve some of the world's most complex problems. For example, AI-powered medical diagnosis systems can provide accurate and fast diagnoses that can save lives. AI can also help scientists analyze large datasets to make new discoveries and breakthroughs. In addition, AI can optimize energy consumption, reduce traffic congestion, and improve transportation efficiency, which can have a significant impact on the environment. Groups like OpenAI are then obligated to address these concerns. It's essential to ensure that AI is developed and deployed responsibly. This means creating guidelines and regulations that prioritize transparency, accountability, and ethical considerations. It also means investing in education and training programs that prepare individuals for the jobs of the future.

AI can also increase efficiency and productivity in various industries. AI-powered systems can automate repetitive and mundane tasks, allowing humans to focus on more complex and creative work. This can lead to increased productivity, faster turnaround times, and improved customer service. For example, chatbots can answer customer inquiries and resolve issues, freeing up human customer service representatives from handling more complex issues.

AI is also essential for businesses which desire to remain competitive in today's digital world. AI-powered tools can help companies analyze customer data, predict consumer behavior, and develop targeted marketing strategies. This can lead to increased sales, customer satisfaction, and loyalty.

However, despite the numerous benefits of AI, there are also concerns about its potential negative effects. For example, some worry AI could lead to job loss as automation replaces human workers. There are also concerns about AI being used for malicious purposes, such as cyberattacks or surveillance.

## Expert Opinions

The public at large has been interested in AI arguably prior to the release of the *Erewhon* novel by Samuel Butler in 1872 or the introduction of data in the original *Star Trek* series, but what are academics' opinions on AI presently and its potential? Geoffrey Hinton (as well as several other big name AI developers) has left his job at Google because of his concerns about AI's rapid, unregulated development. He reminds the public in his BCC interview that AI systems that are currently open to the public are still very much in their Beta stages and are severely limited to what they can understand.

Understanding, in this instance, would refer to a system's ability to response to its text prompts intelligently, consistently, and accurately, a vast majority of the time rather than having human level comprehension (although these systems aren't currently capable of these either by lacking human level autonomy). Hinton has emphasized the need for AI systems to be developed with careful consideration of their potential impacts. He has warned that AI systems can have unintended consequences, especially if they are not designed with sufficient attention to their potential risks.

While groups such as OpenAI, Alphabet, and other conglomerates have admitted to potential flaws and claim to be as transparent as they can without being a security risk to either corporate entities, government offices, or private citizens, Hinton claims that this is still not transparent enough for the information spewed from LLMs to be completely trusted without heavy cross referencing. It still might not even be entirely possible for generative answers to be independently verified depending on the phrasing or because of the citations coming from journals and or websites which aren't available for free public access.

Hinton also alludes to the human oversite in many of these systems are lacking, because there are too few human operatives, or the human operatives are being paid pennies on the dollar in foreign countries which leads to faulty corrections.

Recently, however, there was a conference at the World Economic Forum (WEF) which included Microsoft's corporate VP and chief economist, Michael Schwarz who disagrees with Hinton. Schwarz stated "I am quite confident that yes, AI will be used by bad actors; and yes, it will cause real damage; and yes, we have to be very careful and very vigilant […] we shouldn't regulate AI until we see some meaningful harm that is actually happening – not imaginary scenarios" (Michael Schwarz, WEF Growth Summit, 2023). When questioned, Schwarz claims that problems with AI could not be accurately predicted, even making an analogy to driver's licenses being issued only after there were significant deaths. It's important to note that humans are not incapable of accurately predicting future problems.

Even in the example he gives with cars, people, including experts, were aware of the dangers proposed by unlicensed drivers. Unlicensed drivers allowed large car corporations, like Ford, to sell vehicles without the upfront expenses of regulation. This then allowed companies to lobby for civic planning to make new, car dependent cities. Perhaps then, Mr. Schwarz's is far morre concerned with profit margins than innovation.

He fails to mention a more obvious problem to consumers would be that more regulations means more potential privacy breaches and taxes to rise to accommodate for the new monitoring systems. This too might be an over-exaggeration since the United States government and independent corporations like Microsoft already has access to private citizens' data which can be legally acquired if deemed necessary. It is then foolish to state that significant harm is only hypothetical – it is already here in the form of the mental atrophy of students.

Another critique of AGI is Eliezer Yudkowsky, an artificial intelligence researcher at Berkley. Yudkowsky argues that AI systems have the potential to become super-intelligent and vastly exceed human intelligence. This means that they could outsmart humans in any field, including the design and improvement of their own systems, leading to an intelligence explosion and an exponential increase in AI capabilities. He argues that this could cause a scenario where the AI system's goals diverge from those of its creators and cause unintended consequences that could pose a threat to human existence.

Yudkowsky advocates AGI development safety research, with a focus on creating "friendly AI" that will be aligned with human values. He also advocates for the creation of regulatory frameworks to ensure that AI development is done safely and in the interests of humanity. He believes that the risk posed by advanced AI is significant enough to warrant careful consideration and regulation, and that the potential benefits of AI should be weighed against the potential risks. Recently, he has critiqued the letter signed by notable figures Elon Musk, Steve Wozniak, and Gary Marcus because Yudkowsky believes a 6-month pause is not enough time to create enough ethical AGI systems. Yudkowsky calls for an indefinite worldwide moratorium on large AGI learning operations by imposing limits on LMM systems used for AI training and shutting down large GPU clusters.

## Current Legislation & Future Regulation

Currently, there are many countries outside of the United States who've regulated AI systems like ChatGPT. The European Commission has taken legislative measures to provide broad, but temporary, regulations to AI development, personal and company use, and information gathering techniques. The regulation defines AI as software that is developed with one or more machine learning or other techniques, and can, at least to some extent, learn from data or generate outputs based on data. The definition covers both weak and AGI.

Regulations currently being proposed require specific requirements for certain uses of AI, such as biometric identification, remote biometric identification, critical infrastructure, and educational and vocational training. For example, the use of AI for biometric identification will be prohibited in public spaces, except in certain limited circumstances. Systems would also be banned from using information that would aid in on the books crimes such as terrorism or kidnapping but also in minor instances such as finding private citizens' locations or stealing produce from copyright holders, like pirating a new video game or movies.

The regulation proposes a conformity assessment process for high-risk AI systems, which would involve a third-party assessment of the system's compliance with the regulation. Non-compliant systems could face fines of up to 6% of their global turnover. The commissions' regulation emphasizes the importance of governance and transparency in the development and use of AI systems. AI developers and users must ensure that their systems are transparent, accountable, and respect human rights and fundamental values. The regulation encourages the use of self-regulatory codes of conduct and certification schemes to promote best practices and foster trust in AI systems despite companies or private groups like OpenAI claiming to already practice these virtues.

Circling back to Yudkowsky's proposed regulations, he suggests GPUs and AI/AGI hardware be tracked (by both government bodies and independent organizations) and lowering the computing power cap for AI training. His suggested regulations also disallow government use of AI, claiming the process of data harvesting with insufficiently regulated AI systems be far more profitable than any physical resource currently available – from diamonds, to petrol, to real estate. In an increasingly digital world where everything from learning, to banking, to shopping is online, lives could be stolen or manipulated in picoseconds by a sufficiently competent AGI.

## Conclusion

To conclude, the development of AGI technology poses a potential danger to academia. While the concept of creating a machine that can simulate human-level intelligence is fascinating, the potential consequences cannot be ignored. Just as calculators have been detrimental to students who lack basic math skills, AGI could be detrimental to individuals who lack the necessary cognitive and emotional skills to navigate complex decisions in their personal and professional lives. The development of AGI technology should not be abandoned, but it must be approached with caution and ethical considerations. It is important to ensure that we do not lose sight of the importance of human decision-making and problem-solving skills in our pursuit of technological advancement. The relationship between AGI and the human mind and body is still an unresolved matter that will need more study and discussion down the road. Perhaps then we ought not build the torment nexus.

# Bibliography

Stanford Encyclopedia of Philosophy. (2019). Artificial Intelligence. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2019 Edition). Retrieved from https://plato.stanford.edu/entries/artificial-intelligence/#WhatExacAI

European Commission. (2021, April 21). European approach to artificial intelligence. European Commission - Digital Strategy. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

Sajid, H. (2023, April 29). What are LLM Hallucinations? Causes, Ethical Concerns, and Prevention. Unite.AI. Retrieved from https://www.unite.ai/what-are-llm-hallucinations-causes-ethical-concern-prevention/

Boiko, D. A., MacKnight, R., & Gomes, G. (2022). Emergent Autonomous Scientific Research Capabilities of Large Language Models. arXiv preprint arXiv:2201.03187. Retrieved from https://arxiv.org/abs/2201.03187

United States Copyright Office. (2022). Zarya of the Dawn. United States Copyright Office. Retrieved from https://copyright.gov/docs/zarya-of-the-dawn.pdf

Peijie Jiang "Virtual graphing calculator applied in learning and the performance based on teaching experiment method", Proc. SPIE 12506, Third International Conference on Computer Science and Communication Technology (ICCSCT 2022), 1250641 (28 December 2022); https://doi.org/10.1117/12.2661766

Kokalitcheva, K. (2022, December 7). Google isn't launching a ChatGPT competitor due to 'reputational risk,' according to a report. Business Insider. Retrieved from https://www.businessinsider.com/google-isnt-launching-chatgpt-competitor-due-to-reputational-risk-2022-12

Cope, A., & Irwin, R. (2022). The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions? Issues in Accounting Education. Advance online publication. https://doi.org/10.2308/ISSUES-2023-013

Dwoskin, E. (2023, April 9). OpenAI's safety team is growing as AI-powered tech advances. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2023/04/09/ai-safety-openai/

Vallance, Z. K. & C. (2023, May 2). *Ai "godfather" Geoffrey Hinton warns of dangers as he quits Google*. BBC News. https://www.bbc.com/news/world-us-canada-65452940